MDD Cup 2017 算法大赛赛题

创建： 李腾，最新修改： 侯俊杰 于 十月 01, 2017

1. 赛题背景

在外卖配送场景中，"预计送达时间"是用户体验的重要指标，同时也是调度系统最重要的约束条件。作为每天影响千万级别订单的模块，其重要性不言而喻。
外卖订单生命周期长、经历环节多、外界因素复杂，这些都会影响订单的配送，使"时间预估问题(ETA)"成为整个外卖行业的难题。

订单的生命周期

影响订单送达时长的因素

- 订单：下单时刻、价格、菜品数
- 用户：用户位置
- 骑手：骑手负载
- 商家：出餐能力、地理位置、积压的订单
- 区域：骑手数、单量
- 时间：午高峰期、晚高峰期、非高峰期
- 天气数据：温度、风速、降水量

2. 任务

根据训练数据进行建模，预估测试样本集中订单的送达时长

3. 评价指标

利用真实值与预测值的Mean Absolute Error(MAE)作为评价指标，公式如下：

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N}|y_i - f(x_i)|$$

其中：y_i为真实值，$f(x_i)$为预测值，N为样本数量

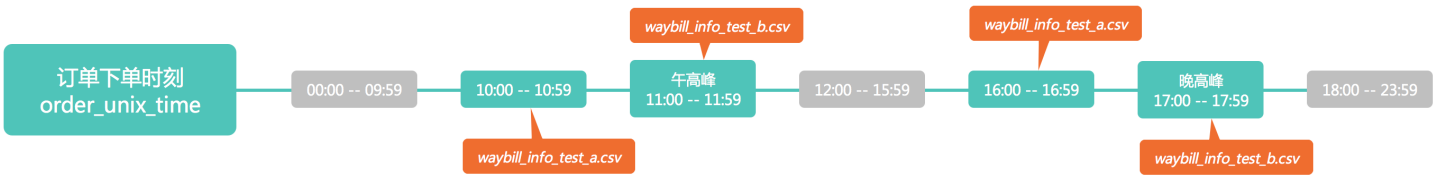
4. 数据集

数据说明

1.订单信息:

(1)训练数据 waybill_info.csv

(2)测试数据



(a) waybill_info_test_a.csv，**用于辅助生产特征，**包含下单时刻(order_unix_time)位于 10:00-11:00 和 16:00-17:00 的订单。

(b) waybill_info_test_b.csv，**用于最终评估，**在kaggle中上传该数据集中订单的预估结果。文件中包含了下单时刻(order_unix_time)位于午高峰(11:00 - 12:00)和晚高峰(17:00-18:00)的订单。

字段	类型	解释	备注
order_id	string	订单id	
poi_id	int	商家id	
area_id	int	配送区域id	

字段	类型	解释	备注
food_total_value	double	订单价格(原价)	
box_total_value	double	餐盒费	
food_num	int	菜品数量	
delivery_distance	double	配送导航距离（米）	
order_unix_time	bigint	用户下单时间戳（unix timestamp），单位"秒"	
arriveshop_unix_time	bigint	骑手到店时间戳（unix timestamp），单位"秒"	waybill_info_test_b.csv中不存在此列
fetch_unix_time	bigint	骑手取餐时间戳（unix timestamp），单位"秒"	waybill_info_test_b.csv中不存在此列
finish_unix_time	bigint	骑手送达时间戳（unix timestamp），单位"秒"	waybill_info_test_b.csv中不存在此列
customer_longitude	double	用户经度	
customer_latitude	double	用户纬度	
poi_lng	double	商户经度	
poi_lat	double	商户纬度	
waiting_order_num	int	下单时刻商户未完成单量	
delivery_duration	int	delivery_duration = finish_unix_time - order_unix_time，单位为"秒"	waybill_info_test_b.csv中不存在此列

2.区域实时特征:

(1)area_realtime.csv:训练数据

(2)area_realtime_test.csv:测试数据

字段	类型	解释	备注
date	string	记录日期 yyyyMMdd	
time	string	记录时间 hhmm	
log_unix_time	bigint	记录时间戳（unix timestamp），单位"秒"。每分钟记录一次	
area_id	int	配送区域id	
working_rider_num	int	区域在岗骑手数(包含忙碌骑手)	
notbusy_working_rider_num	int	区域在岗骑手数(排除忙碌骑手)	
not_fetched_order_num	int	区域未取餐单量	
deliverying_order_num	int	区域取餐未送达单量	

3.天气实时特征:

(1)weather_realtime.csv:训练数据

(2)weather_realtime_test.csv:测试数据

字段	类型	解释	备注
date	string	记录日期 yyyyMMdd	
time	string	记录时间 hhmm	
log_unix_time	bigint	记录时间戳（unix timestamp），单位"秒"。	
area_id	int	配送区域id	
temperature	double	温度	
wind	double	风速	
rain	double	雨量	

预估目标：订单送达时长 delivery_duration = finish_unix_time - order_unix_time，单位为"秒"

5.要求

- 提示：请选手注意**测试数据的使用格式**，以免基于训练数据构造的特征无法在测试中使用。

- 参赛者上传的预估结果需要包含两列: 订单id(order_id), 预计送达时长(delivery_duration)。**订单id与测试集（waybill_info_test_b.csv）中保持一致**，预计送达时长单位为”秒“。

  数据保留表头, 以逗号分割，格式如下

  order_id, delivery_duration

  1503590849160934,1800

  1503590849160935,1800

  1503590849160936,1800

  1503590849160937,1800

  1503590849160938,1800

  .....

- 进入决赛的参赛者需要**上传完整代码**，包括：数据预处理、模型训练等，接受命题组的code review

返回大赛综述页面

赞 11人赞了它

无标签